http://www.math.nyu.edu/faculty/goodman/teaching/MonteCarlo20/

Class notes: Monte Carlo methods
Week 13, theory
tentative Jonathan Goodman
Tentative December 9, 2020

1 Convergence rates

Suppose X_n are the steps of a Markov chain and $X_n \sim \rho_n$ are the corresponding PDFs. Under some non-degeneracy conditions, we know that $\rho_n \to \rho$ as $n \to \infty$, with ρ being the target invariant distribution. We have seen that ρ_n can converge to ρ slowly. There are theoretical approaches that lead to proofs that certain MCMC algorithms converge at certain rates.

To prove that ρ_n is close to ρ, you need a notion of distance between probability distributions. There are many different ways to measure distance. They give similar answers for simple distributions in low dimension, but remarkably different answers for complex distributions in high dimensions.

Total variation

Total variation is one measure of the difference between probability distributions. Suppose $\rho(x)$ and $\sigma(x)$ are two probability densities in $\mathbb{R}^d$. The total variation distance between them is

$$\|\rho - \sigma\|_{\rm TV} = \int_{\mathbb{R}^d} |\rho(x) - \sigma(x)| \, dx \; . \tag{1}$$

This is the same as the L^1 norm, as long as you're talking about probability densities. There is a more abstract definition of total variation distance that is useful in more abstract situations. We define the ρ (or σ) probabilitiy of a set A by

$$\rho(A) = \Pr_\rho(X \in A) = \int_A \rho(x) \, dx \; .$$

This definition make contact between probability densities and probability measures. A probability measure is any way of defining probabilities $\rho(A)$, which may not involve integrating a probability density. The probability measure definition of total variation difference between probability measures is

$$\|\rho - \sigma\|_{\rm TVm} = 2 \sup_A \left[\rho(A) - \sigma(A)\right] \; . \tag{2}$$

If $\rho(A) = \sigma(A)$ for all A, then this gives zero, as it should. Otherwise, there is some with $\rho(A) \neq \sigma(A)$. If $\rho(A) > \sigma(A)$, this shows $\|\rho - \sigma\|_{\rm TVm} > 0$. If $\rho(A) < \sigma(A)$, then $\rho(A^c) > \sigma(A^c)$ so again $\|\rho - \sigma\|_{\rm TVm} > 0$. [In this, "sup" is for *supremum*. If you haven't taken a technical analysis class, this is slightly different from "max" for *maximum* in that the *supremum* need not be

"achieved". For example, the sequence $a_n = 1 - \frac{1}{n}$ has supremum equal to 1 but it has no maximum because $a_n < 1$ for all n. If you have taken a technical analysis class, you could prove in this case (using "countable additivity") that there is a set A that achieves the supremum, so it could have said "max".] If the measures ρ and σ are given by probability densities, the definitions are the same. We prove this using the "two inequalities" method. You can show that $Q_1 = Q_2$ by showing that $Q_1 \le Q_2$ and $Q_1 \ge Q_2$.

First, we prove that $\|\rho - \sigma\|_{\mathrm{TVm}} \ge \|\rho - \sigma\|_{\mathrm{TV}} > 0$. For this, take $A = \{x \mid \rho(x) > \sigma(x)\}$. Then, using the fact that $\Pr(A^c) = 1 - \Pr(A)$, so $\rho(A) = 1 - \rho(A^c)$, etc.,

$$\begin{aligned}
\|\rho - \sigma\|_{\mathrm{TV}} &= \int_{\mathbb{R}^d} |\rho(x) - \sigma(x)| \, dx \\
&= \int_A [\rho(x) - \sigma(x)] \, dx - \int_{A^c} [\rho(x) - \sigma(x)] \, dx \\
&= \rho(A) - \sigma(A) - [\rho(A^c) - \sigma(A^c)] \\
&= \rho(A) - \sigma(A) - [1 - \rho(A) - \{1 - \sigma(A)\}] \\
&= 2\,[\rho(A) - \sigma(A)] \\
&\le \|\rho - \sigma\|_{\mathrm{TVm}} \ .
\end{aligned}$$

In the last line, we used the fact that $\sup_A Q(A) \ge Q(A)$ for any specific A.

The proof in the other direction uses some of the same algebra. Suppose A_* is the A that achieves $\|\rho - \sigma\|_{\mathrm{TVm}}$ in (2).

$$\begin{aligned}
2\,\|\rho - \sigma\|_{\mathrm{TVm}} &= 2\,\{\rho(A_*) - \sigma(A_*)\} \\
&= \int_{A_*} [\rho(x) - \sigma(x)]\, dx + \int_{A_*^c} [\sigma(x) - \rho(x)]\, dx \\
&\le \int_{A_*} |\rho(x) - \sigma(x)|\, dx + \int_{A_*^c} |\rho(x) - \sigma(x)|\, dx \\
&= \|\rho - \sigma\|_{\mathrm{TV}} \ .
\end{aligned}$$

A third version of total variation distance involves difference in expected values. The supremum here is over all functions $f(x)$ with $|f(x)| \le 1$ for all x.

$$\begin{aligned}
\|\rho - \sigma\|_{\mathrm{TVf}} &= \sup_f \left[\mathrm{E}_\rho[f(X)] - \mathrm{E}_\sigma[f(X)]\right] \qquad (3) \\
&= \sup_f \left[\int_{\mathbb{R}^d} f(x)\rho(x)\, dx - \int_{\mathbb{R}^d} f(x)\sigma(x)\, dx\right] \ .
\end{aligned}$$

You can prove that this definition is the same as $\|\cdot\|_{\mathrm{TV}}$ by taking

$$f(x) = 1 \ \text{ if } \ \rho(x) > \sigma(x) \ , \ \text{ and } \ f(x) = -1 \ \text{ if } \ \rho(x) < \sigma(x) \ . \qquad (4)$$

You get some insight into the use of this definition by considering an inequality that follows from it

$$|\,\mathrm{E}_\rho[f(X)] - \mathrm{E}_\sigma[f(X)]\,| \le \sup_x |f(x)|\, \|\rho - \sigma\|_{\mathrm{TV}} \ .$$

If two probability distributions are close in the total variation sense, then they give similar expected values for any bounded function, even if it's discontinuous as (4).

The term "total variation" is traditional but not completely appropriate. It arose from the sensible definition of *total variation* of a function $f(x)$ of one variable, as

$$\mathrm{TV}(f) = \int |f'(x)|\,dx \ .$$

If f is not differentiable, the definition is

$$\mathrm{TV}(f) = \sup_n \sup_{x_1<\cdots<x_n} \sum_{k=1}^{n-1} |f(x_{k+1}) - f(x_k)| \ .$$

This total variation is equal to an L^1 integral of f', which is how the term "total variation" came to be associated with L^1.

Wasserstein distance

The *Wasserstein* distance is harder to explain but very helpful in high dimensional applications. It starts with the idea of a *coupling* between probability measures. Suppose $\rho(x)$ and $\sigma(x)$ are probability densities on $\mathbb{R}^d$. Suppose $p(x,y)$ is a probability on $\mathbb{R}^{2d} = \mathbb{R}^d \times \mathbb{R}^d$. We say that p is a *coupling* between ρ and σ if

$$(X,Y) \sim p \implies X \sim \rho \ \text{ and } \ Y \sim \sigma \ .$$

The set of all couplings is (according to Wikipedia, which is always right) $\Gamma(\rho,\sigma)$.

Couplings are used in Monte Carlo theory, as today, and also in Monte Carlo practice, as last week with the Brownian bridge construction. A coupling is a "story" that gives a way X and Y could have been generated at the same time but with $X \sim \rho$ and $Y \sim \sigma$. There is always the "trivial" coupling in which X and Y are independent. This has $p(x,y) = \rho(x)\sigma(y)$. If ρ and σ are close, it should be possible to create "better" couplings with X closer to Y.

For example, suppose X and Y are one component random variables with $X \sim \mathcal{N}(0,1)$ and $Y \sim \mathcal{N}(\mu,1)$. You can couple X to Y by taking $Y = X + \mu$. In probability densities, this is $p(x,y) = \rho(x)\delta(y-x)$. Similarly, if $Y \sim \mathcal{N}(0,v)$ with $v > 1$ (can't say σ^2 because $\sigma(x)$ is the PDF), then we can take $Y = X + Z$ where $Z \sim \mathcal{N}(0, v-1)$ and Z independent of X.

The Wasserstein distance involves the expected distance between X and Y, in the best coupling:

$$W_p(\rho,\sigma) = \inf_{\Gamma} \ \mathrm{E}[\,|X-Y|^p]^{\frac{1}{p}} \ . \tag{5}$$

The exponent p should be between 1 and ∞, with $p = 1$ and $p = 2$ being the most common. This definition seems harder to apply that total variation (1) because it involves more than just integration. Nevertheless, it is useful in some proofs where natural couplings can be constructed. We seek Wasserstein differences for complex problems, particularly in high dimensions, where total variation is too strong.

2 Lyapunov functions – control the tails

A Markov chain on an infinite state space might "wander to infinity". This is a way to avoid having an invariant distribution. Random walk in one dimension does this, for example. On the other hand, consider a random walk in one dimension with $X \geq 0$ (rejecting proposals that go to $X < 0$ and "drift" toward 0. For example, $X_{n+1} = X_n - \mathcal{N}(-\mu, \sigma^2)$ with $\mu > 0$, rejecting proposals $X_{n+1} < 0$.

3 Cheeger inequality – control bottlenecks

An MCMC method can be slow if it has *bottlenecks*. A bottleneck is an unlikely set that X_n has to pass through to get from one likely set to another. The double well potential with equal depth wells has a bottleneck at the local maximum of the potential between the two wells. The MCMC process will spend a long time on one well before making a transition to the other.

The *conductance* is a way to define and measure bottlenecks. A set A defines a bottleneck if $\rho(A) \leq \frac{1}{2}$ but if $X_n \in A$ then $X_{n+1} \in A$ with high probability. The technical definition of conductance is

$$\Phi = \inf_{\rho(A) \leq \frac{1}{2}} \Pr_\rho(X_{n+1} \notin A \mid X_n \in A) \tag{6}$$

We need some restriction like $\rho(A) \leq \frac{1}{2}$, though not this one exactly, because otherwise we could take A to be the whole space and get $\Phi = 0$ because it's impossible to escape the whole space.

Jeff Cheeger showed (in a different context but with the same ideas and almost the same definition) that the process has a *spectral gap* of size at least Φ^2. There is a beautiful inequality of Lovasz and Simonovitz that gets to Cheeger's point in a different way. This goes by a function $u_n(\lambda)$ defined for $0 \leq \lambda \leq 1$ defined by

$$u_n(\lambda) = \sup_{\rho(A) = \lambda} \rho_n(A) - \lambda \; .$$

If $\rho_n = \rho$ then $u_n(\lambda) = 0$ for all λ. It is not hard to show that $u_n(\lambda)$ is a concave function of λ. The inequality is

$$u_{n+1}(\lambda) \leq \frac{1}{2} \left[u_n((1 - \Phi^2)\lambda) + u_n((1 + \Phi^2)\lambda) \right] \; . \tag{7}$$

This shows that $u_{n+1}(\lambda) \leq u_n(\lambda)$ for all λ. It also shows (with more work) that $u_n \to 0$ as $n \to \infty$ at a rate that depends on $1 - \Phi^2$. This gives the same convergence rate as Cheeger's spectral gap, but in a stronger sense.